

日英機械翻訳システム TWINTRAN の 言語知識と翻訳品質の評価

吉見 毅彦[†] Jiri Jelinek^{††} 西田 収^{†††}
田村 直之^{††††} 村上 温夫^{†††††}

本稿では、日英機械翻訳システム TWINTRAN の言語知識 (辞書と規則) と翻訳品質の評価結果について述べる。TWINTRAN は次のような設計方針に基づくシステムである。1) 翻訳対象と翻訳方向を日本語テキストから英語テキストへの翻訳に固定し、日本語の解析知識の記述を、日本語文法だけでなく英語文法も考慮に入れて行なう。2) 解釈の曖昧性の解消は、各規則に与えた優先度に基づいて解釈の候補に優劣を付け、候補の中から最も優先度の高い解釈を選択することによって行なう。3) 動詞の主語や目的語など日本語では任意的だが英語では義務的である情報を得るために照応解析を行なう。

NTT 機械翻訳機能試験文集を対象として行なったウィンドウテストでは、我々の評価基準で、試験文集の 73.1% の文が合格となり、試験文集全体の平均点も合格点を上回る結果が得られた。

キーワード: 機械翻訳, 曖昧性解消, 優先度, 照応解析, 日英対照分析

Linguistic Expertise in the Japanese-to-English MT System TWINTRAN and Evaluation of Translation Results

TAKEHIKO YOSHIMI[†], JIRI JELINEK^{††}, OSAMU NISHIDA^{†††},
NAOYUKI TAMURA^{††††} and HARUO MURAKAMI^{†††††}

In this treatise we set out to describe the linguistic expertise (dictionary and rules) embodied in the Japanese-to-English MT system TWINTRAN, and the evaluation of the translation results. TWINTRAN is based on the following design policy: 1) The translation equivalents and the direction of the translation process are strictly monodirectional, from Japanese to English. The analysis of the Japanese input is not confined to Japanese grammar but also anticipates at every step the possible English translation. 2) Disambiguation is based on prioritisation of each rule, where each rule contains a priority value and the highest aggregate priority candidate is selected. 3) Verb complements are screened for acceptability not only in the input Japanese but also in the output English, and anaphora resolution is used for arriving at the optimum result.

In the window test we have carried out, based on NTT's functional MT test set, applying our evaluation procedure, 73.1% of the corpus was acceptable and the corpus average was above the point of acceptability.

KeyWords: *Machine Translation, Disambiguation, Prioritisation, Correferentiality Analysis, Japanese-to-English Contrastive Analysis*

1 はじめに

これまでに開発されている機械翻訳システムの多くはトランスファ方式に基づいており、原言語の性質だけに依存する解析辞書・規則と、原言語と目的言語の両方の性質に依存する対照辞書・規則が個別に記述されている。他方、翻訳対象言語対と翻訳方向を固定した上で、解析知識の記述を、原言語の性質だけでなく目的言語の性質も考慮に入れて行なうという設計方針もある。このような方針を採ると、ある原言語（例えば日本語）の解析知識を、異なる目的言語へのシステム（例えば日英システムと日中システム）で共用できるというトランスファ方式の利点が失われ、別の目的言語へのシステムを開発する場合には新たな解析知識の記述が必要になる。しかし、当面、ある特定の原言語から特定の目的言語への翻訳に焦点を絞れば、以下ののような利点が得られる。

- (1) 原言語の表現と目的言語の表現を比較的表層のレベルで対応付けることができる。例えば動詞と補足語との結合関係は、翻訳対象と翻訳方向を日本語テキストから英語テキストへの翻訳に固定した場合、日本語の助詞と英語の前置詞との対応として記述できる。従って、結合関係を深層格として抽象化する必要がなくなり、深層格の認定基準の設定などの困難が避けられる。
- (2) 翻訳対象言語対と翻訳方向を固定しない場合、ある目的言語の生成には必要ないが別の目的言語の生成には必要な情報も解析過程で抽出しておく必要がある。これに対して、翻訳対象と翻訳方向を固定すると、原言語を解析する知識を記述する際に目的言語の性質を考慮することが可能になるため、目的言語の生成に必要な情報のみを抽出すればよくなり、無駄な処理が避けられる。

このようなことから、我々のシステム TWINTRAN では、目的を日英翻訳に限定した上で、英語の適切なテキストを生成するためには日本語テキストがどのように解析されていなければならないかという観点から辞書と全規則群を記述している。辞書では、英語に翻訳する際にこれ以上分解するとその意味が変化してしまう表現はそれ以上分解せずに一見出しとして登録する(2節)。構文解析規則では、動詞と補足語との結合関係を日本語の助詞とその英訳との対応に基づいて区別し、動詞型を日本語の結合価パターンと英語の結合価パターンとの対に基づいて設定する(4節)。また、日本語の連体従属節を英語の関係節に翻訳するための関係詞決定規則を設ける(5節)。日本語で明示することは希であるが英語では明示しなければならない言語形式上の必須情報(名詞句の定/不定性の区別や、動詞の主語や目的語になる代名詞など)を得るために、照応解析規則を陳述縮約パラダイム(Jelinek, Yoshimi, Nishida, Tamura and Murakami

† シャープ(株) 情報システム事業本部, Information Systems Group, SHARP Corporation

†† 所属先なし, None

††† シャープ(株) 情報家電開発本部, Digital Information Appliances Development Group, SHARP Corporation

†††† 神戸大学 工学部 情報知能工学科, Department of Computer and Systems Engineering, Faculty of Engineering, Kobe University

††††† 神戸大学名誉教授, Emeritis Professor of Kobe University

1995) に基づいて記述する (6.2 節).

TWINTRAN と同じように日本語の解析知識を日英対照の観点から記述しているシステムとして, ALT-J/E(Ikehara, Shirai and Bond 1996; 中岩, F. 1997) や, US 式翻訳システム (柴田 1996, 1998) などがこれまでに報告されている. これらのシステムでも照応解析が行なわれているが, 構造を持たない言語表現の間で成立する照応すなわち語と語の間の照応の解析に留まっている. これに対して TWINTRAN では, 文や句など構造を持つ言語表現間の照応も扱う¹.

機械翻訳システムにおける重要な課題の一つは, テキスト解釈の曖昧性を解消し妥当な解釈を一意に決定することである. 曖昧性解消へのアプローチには, 言語知識を絶対的な基準 (制約) とみなす立場と, 相対的な比較基準 (選好) とみなす立場がある (長尾, 丸山 1992). 前者では, ある解釈を受理するか棄却するかの判断は, その解釈と他の解釈を比較せずに行なわれる. これに対して後者では, ある解釈の選択は他の解釈との比較に基づいて行なわれる (Wilks 1978; Tsujii, Muto, Ikeda and Nagao 1988; 島津, 内藤 1989; Hobbs and Bear 1990; Den 1996). TWINTRAN では後者の立場から, 各規則に優先度を付与し, それに基づいて解釈の候補に優劣を付け, 候補の中から最も優先度の高い解釈を選択することによって曖昧性の解消を行なう. テキスト解析では, 形態素解析規則から照応解析規則に至るまでいくつかの種類の規則が利用され, それぞれ異なる観点から一つの解釈の良さが評価される. このとき, ある種類の規則による解釈の良さと他の種類の規則による解釈の良さが競合する可能性があるため, 各観点からの評価をどのように調整するかが重要となる. TWINTRAN では, 構文, 共起的意味, 照応に関する各規則による優先度の重み付き総和が最も高い解釈をテキストの最良解釈とする (6.3 節).

以下, 2 節ないし 9 節で各処理過程について説明し, 10 節で翻訳品質評価実験の結果を示す.

2 辞書

TWINTRAN には, 解析に必要な情報だけでなく変換や生成に必要な情報の大半も記述した辞書と, 屈折形や派生形の生成に必要な情報を記述した辞書が存在する. 以降, 前者を単に辞書と呼ぶ. 辞書には現在約 13 万見出しが登録されており, 各見出し表現に与えられている情報は形態素前接番号, 形態素後接番号, 同形異義に関する優先度 (自然数), 語種 (品詞), 人称, 性, 数, 意味標識, 共起制約, 陳述縮約度, 対訳情報である. 各情報についての説明は, それらを利用した処理について述べる節で行なう.

辞書は, 英語でのひとまとまりの表現に対応する日本語でのひとまとまりの表現を認識することを目的としている. このため, 日本語文法単独ではひとまとまりの表現ではないが日英翻訳の観点からはひとまとまりとみなすべき表現は一つの見出しとして登録している². 例として,

1 より精度の高い解析を実現するためには, 構造を持つ表現の意味をその部分から導き出す必要があるが, その実現は今後の課題である.

2 従って, 形態素解析でひとまとまりと認識される単位は, 通常の単語あるいは文節とは必ずしも一致しない.

「雨」という語で始まる見出しの一部について語種と対訳情報を表 1 に示す。

語種 “sgnpli” と “ntmo” と “nuc” は名詞類, “a2v” は副詞類, “vel” は動詞類, “subj” は主語格標識をそれぞれ意味する。「雨が降(る)」と「雨模様(だ)」には, 一つの語種ではなく “nuc subj vel” という語種列を与えているが, これによって一つの見出しを構文木上の複数の終端節点に対応付けている. この語種列からどのような構文木が構成されるかは 4 節で述べる.

実際の辞書の対訳情報には, 英単語だけでなく, 変換と生成のための特殊記号が含まれているが, 表 1 では簡単のために英単語のみを示した. 実際の処理では, この特殊記号に基づいて動詞の屈折形などを決定する.

表 1 辞書見出しの一部

見出し表現	語種	対訳情報
雨	sgnpli	rain
雨の日	ntmo	rainy day
雨が降ろうと槍が降ろうと	a2v	come hell or high water
雨後の筍のように乱立	vel	mushroom
雨が降	nuc subj vel	it rain
雨模様	nuc subj vel	it look like rain

3 形態素解析

形態素解析では, 辞書情報のうち見出し表現, 形態素前接番号, 形態素後接番号, 同形異義に関する優先度を参照する. さらに, 後接番号と前接番号の接続可能性を記述した接続表を利用する. 接続表には, 接続可能な後接番号と前接番号の組と, その接続に関する優先度 (0.1, 0.2, ..., 1.0 の十段階) が記述されている.

形態素解析は, チャート構文解析法 (Kay 1980) の原理に従って行なう³. 入力文を構成する各文字の間に番号付きの節点が存在すると考えると, チャート法による形態素解析は, 前述の前接, 後接番号と接続表に基づいて, 節点の間を結ぶ弧を順次生成することによって進む. $\beta_1 \cdots \beta_i$ から成る弧 α と, α の右側に隣接する弧 β_{i+1} を結合して新たな弧 α' を生成するとき, 弧 α' に与える優先度は 1) 弧 α が持つ接続優先度と, β_i と β_{i+1} の接続優先度との積と, 2) 弧 α が持つ同形異義優先度と, β_{i+1} の同形異義優先度との和の二種類とする.

形態素解析での最良解釈の選択は, これら二種類の優先度に基づいて二段階で行なう. まず, ある同一節点区間を結ぶ弧の中から, 最も高い接続優先度を持つ弧を選び出し, それ以外の弧は棄却する. この選択で解釈が唯一に決まれば, それを最良解釈とする. さもないければ, 接続優先度が最も高い弧の中から同形異義優先度が最も高い弧を選び, それ以外は棄却する. 同形

3 もちろん構文解析と異なり, 階層的な木構造は生成しない.

異義優先度が最も高い弧が複数存在する場合、それらすべてを最良解釈とする。

形態素解析の後処理として、構文的な構造を作る必要がない語種、すなわち構文木の一節点とみなす必要がない語種を一つにまとめる。例えば動詞の語幹を表す語種と屈折を表す語種をまとめる処理を行なう。

4 構文解析

拡張文脈自由文法形式に基づいて、構文解析規則の記述形式を次のように定める。

$$A ::= B_1 \cdots B_i / \text{head} \cdots B_m, \{Aug\}, Prio, DL \quad (1)$$

ここで、 B_i / head は B_i の主辞が A の主辞になることを意味し、 $\{Aug\}$ は補強項であり、 $Prio$ はこの規則の適用優先度 (自然数)、 DL は構文範疇 A の陳述縮約度 (Jelinek et al. 1995) である。陳述縮約度については 6.2 節で説明する。

TWINTRAN に実装されている構文解析規則の総数は 1734 であり、規則記述に用いられている終端構文範疇の数 (異なり数) と非終端構文範疇の数 (異なり数) はそれぞれ 307 と 208 である。規則の一部を図 1 に示す。

(a)	TEXT	::=	SENTENCE/head,	{},	0, 0
(b)	TEXT	::=	SENTENCE TEXT	{},	0, 0
(c)	SENTENCE	::=	THM SENT/head end,	{},	1, 1
(d)	SENTENCE	::=	SENT/head end,	{},	1, 1
(e)	THM	::=	NP/head thm,	{},	4, 6
(f)	SENT	::=	VPAT ve1/head,	{check_vpat(ve1)},	0, 2
(g)	SENT	::=	ve1/head,	{},	0, 2
(h)	VPAT	::=	SUBJ/head,	{},	0, -
(i)	VPAT	::=	SUBJ VPAT,	{},	0, -
(j)	VPAT	::=	OBJ/head,	{},	0, -
(k)	VPAT	::=	OBJ VPAT,	{},	0, -
(l)	SUBJ	::=	NP/head subj,	{},	0, 5
(m)	NP	::=	nuc/head,	{},	0, 7

図 1 構文解析規則

図 1 の規則 (a) と (b) は、複数の文から成るテキスト全体を一つの構文木で表現するための

規則である．規則 (h) ないし (k) は，動詞の補足語をまとめる規則である．動詞と補足語の結合関係は，深層格として抽象化せず，日本語の助詞とその英訳との対応に基づいて表層で区別し，表 2 の 16 種類を設定している．

表 1 の「雨が降(る)」に与えられている語種列 “nuc subj ve1” からは，規則 (m), (l), (h), (f) によって次のような構文木が構成される．

SENT(VPAT(SUBJ(NP(nuc) subj)) ve1)

表 2 動詞と補足語の結合関係の一覧

結合関係	説明
AG	日本語では助詞「に」や「によって」などで表され，英語では前置詞 “by” で表される．可能態の動詞とこの関係で結合している補足語は英語では主語になることがある．例えば「 <u>私には</u> その理由が理解できる。」は “I can understand the reason.” と英訳される．
AI	日本語では助詞「で」や「によって」などで表され，英語では前置詞 “by” や “with” で表される．
CP	日本語では助詞「と」などで表され，英語では前置詞 “with” で表される．ここで “with” は道具や付加ではなく同伴を意味する．
IO	間接目的語．日本語では助詞「に」などで表され，英語では前置詞 “for” や “from” や “to” で表される．
LT	日本語では助詞「まで」などで表され，英語では前置詞 “as far as” や “up to” で表される．
ML	日本語では助詞「までに」などで表され，英語では前置詞 “by” で表される．
OBJ	日本語では助詞「を」や「は」などで表され，英語では直接目的語となる．
PAC	日本語では助詞「で」や「において」などを場所名詞に付加した形で表され，英語では前置詞 “at” や “in” や “on” を場所名詞に付加した形で表される．
PST	日本語では助詞「に」や「において」などを場所名詞に付加した形で表され，英語では前置詞 “at” や “in” や “on” を場所名詞に付加した形で表される．
PTR	日本語では助詞「を」などを場所名詞に付加した形で表され，英語では前置詞 “across” や “over” や “through” を場所名詞に付加した形で表される．
QA	変化を表す動詞によってもたらされる物事や特性を表す．例えば「彼は <u>校長に</u> 昇進した。」は “He has been promoted to a schoolmaster.” と英訳される．
QO	引用．日本語では助詞「と」などで表される．
SOBJ	日本語では助詞「が」などで表され，英語では直接目的語になる．例えば「彼は <u>林檎が</u> 好きだ。」は “He likes apples.” と英訳される．
SUBJ	日本語では助詞「が」や「は」などで表され，英語では主語となる．
TG	目標．日本語では助詞「に」や「へ」などで表され，英語では前置詞 “for” や “to” で表される．
AUTO	時間表現や状況表現などと任意の動詞との間の関係を表す．任意の動詞と結合し得る．

規則 (f) の手続き check_vpat(ve1) は，VPAT としてまとめられた補足語のすべてが動詞型 ve1 の動詞と結合できるかどうかを表 3 に基づいて調べる．表 3 は，動詞がどのような補足語と結合できるかという観点から能動態の動詞を 19 種類に分類したものの一部である．この他に，受動態，使役態，間接受動態，可能態と，これらの組合せ(受動使役態など)について結合可能

な補足語の種類を記述している。

表 3 結合可能な補足語に基づく動詞の分類

動詞型	結合可能な補足語	例
va2	AI, CP, LT, MT, OBJ, PAC, QO, SUBJ, TG	入力する, 理解する, 打ち明ける
vb1	AI, CP, IO, LT, MT, OBJ, PAC, SUBJ	望む, 受け継ぐ, 借りる
vd1	AI, CP, LT, ML, PAC, QA, SUBJ	現れる, 減退する, 重なる
ve1	AI, CP, LT, ML, PAC, TG, SUBJ	引く, 揚がる, 近づく
vg1b	AI, CP, IO, LT, ML, PST, SUBJ	存在する, 迷う, 酔う
vh	AI, CP, LT, MT, PAC, PTR, SUBJ, TG	急ぐ, 飛び越える, 泳ぐ

構文解析規則の記述では, 3節で述べた形態素解析での優先度や, 6節で述べる共起的意味や照応に関する優先度と異なり, 優先的に適用したい規則ほど小さい数値を与える. 各規則に付与されている優先度の分布を表4に示す. A を頂点とする構文木の優先度は, B_i ($1 \leq i \leq m$) を頂点とする構文木の優先度の和に, 規則の優先度 $Prio$ を加えた値とする. 終端構文範疇を頂点とする構文木の優先度は 0 とする.

表 4 構文優先度の分布

優先度	0	1	2	3	4	5	6	7	8	12	20	30
規則数	1545	122	36	1	14	8	2	1	1	2	1	1

構文解析では最良解釈の選択は行なわず, 規則で許されるすべての解釈を出力する. これらは圧縮したチャート表現 (田中 1989) で表される.

5 英語必須補足語の補完と関係節の処理

よく知られているように, 日本語動詞の主語や目的語などはそれらが何を指しているかが文脈から推測できる場合には明示する必要がないのに対して, 英語動詞の主語や目的語は言語形式上の必須要素であるため代名詞などとして明示する必要がある. 従って, 適切な代名詞を補うことが, 品質の高い英語テキストを生成するためには必要不可欠である.

TWINTRAN では, 適切な代名詞を決定する処理を 1) まず, どの補足語を補う必要があるか (例えば主語を補うべきか目的語を補うべきか) を決定し, 2) 次に, 補われた補足語を具体的にどのような代名詞 (I, you, she など) として英訳するかを決定するという二段階で行なう. 後者の処理は 6 節で述べる共起的意味解析と照応解析で行なう. 前者の処理は図 2 に示すような規則に従って実行する. 図 2 の規則は, 現在着目している節の主動詞が vg3 型の能動態である

場合の規則であり、1)SUBJ も AG もその動詞と結合していないならば SUBJ を補い、2)SUBJ は結合していないが AG が結合しているならば AG を SUBJ とみなすことを意味する。

```

if(head_verb(vg3)){
  if(~exist(SUBJ)){
    if(~exist(AG)){
      create(SUBJ);
    }
    else{
      relabel(AG, SUBJ);
    }
  }
}
}

```

図 2 必須補足語補完規則

関係節として英訳しなければならない従属節に補足語を補う場合、どの補足語を補うかの決定は、関係詞の先行詞になる名詞が従属節の主動詞とどのような関係で結合するかを判定した後、すなわち関係詞を決定した後に行なう。関係詞を決定する規則は 1) 関係詞の先行詞になる主名詞の意味標識、2) 従属節の主動詞の型、3) 従属節の主動詞と結合している補足語の種類に基づいて記述されている。

関係詞決定規則には適用の優先度を五段階で付与しており、優先度の高い規則から順に適用を試み、ある段階の優先度の規則の適用が成功すれば、それより低い優先度の規則の適用は行なわない。この優先度は構文に関する優先度に加算される。この優先度も 4 節で述べた構文に関する優先度と同じく、数値が小さいほど良い。図 3 に示す第一の規則は、最も優先度が高い規則群に属する。この規則は、従属節の主動詞が OBJ と結合できる場合の規則であり、関係詞の先行詞になる主名詞の意味標識が HUM(人間) の下位概念を表し、かつ従属節の主動詞に OBJ が結合していなければ、関係詞 “whom” を構文木に補うことを意味する。第二の規則は、他のどの規則も適用できなかった場合に無条件で適用される規則であり、関係詞 “whereby” を生成する。

6 共起的意味・照応解析

共起的意味解析と照応解析は一つの枠組で統合的に実行する。共起的意味・照応解析では、AND/OR グラフ上での遅延評価による優先度計算機構 (Tamura, Bos, Murakami, Nishida, Yoshimi and Jelinek 1991; 吉見, Jelinek, 西田, 田村, 村上 1997) によって、テキスト解釈候補の中から (準) 最良解釈を効率的に選び出す。


```

if(head_verb(va1)|head_verb(va2)|head_verb(vb1)|head_verb(vb2)|
   head_verb(vc)|head_verb(vd2)|head_verb(vg1a)|head_verb(vg3)){
  if(isa(head_noun, sem_cat:HUM)){
    if(~exist(OBJ)){
      select(whom);
      priority(1);
    }
  }
}

if(true){
  select(wherby);
  priority(5);
}

```

図 3 関係詞決定規則

6.1 共起的意味解析

動詞とその補足語の間の共起的意味解析は、上位下位関係を記述した意味体系を参照しながら、動詞の結合価パターン⁴の各スロットに記述されている共起制約と補足語の人称、性、数、意味標識を照合し、その結果に応じて共起的意味に関する優先度を与える。もし必要ならば、共起制約に違反する解釈も生成する⁴。なぜならば、構文、共起的意味、照応の観点からの優先度がそれぞれ最も高い解釈から、システムが持つ知識全体での最良解釈が構成されるとは限らず、共起制約に違反する解釈からシステム全体での最良解釈が生成される可能性もあるからである。

共起制約に違反する解釈を必要に応じて生成するもう一つの理由は、意味的に不適格な文に対して頑健な処理を実現するためである。このような処理が必要になる表現の典型的な例は比喩である。共起制約に違反する解釈を棄却する立場では比喩に対する解釈を生成することができない。これに対して、共起制約に違反する解釈であっても相対的に見て最良解釈であるならば、これを受理する本稿のような立場では、比喩に対しても解釈を生成することができる。ただし、共起制約に違反する解釈をすべて比喩として受理するのではなく、比喩とみなせる解釈とそうでない解釈を弁別する処理が必要である (Wilks 1978; Ferrari 1996) が、TWINTRAN では実現されていない。

共起的意味に関する優先度は、共起制約が満たされる場合 10 とし、満たされない場合には -6 としている。

6.2 照応解析

適格なテキストでは文と文のつながりによって結束が維持されている。テキストの結束を維持する言語的手段には、照応、代用、省略、接続表現の使用、語彙的つながりがある (Halliday

⁴ 遅延評価による処理を行なっているため、優先度が低い解釈は必要がない場合には生成されない。

and Hasan 1976). 照応は、複数の言語表現が一つの事象に言及することによってテキストの結束を生む手段であり、文や句や語などの様々な言語表現の間で成立する。例えば次のテキスト1では、文「ソ連の国家非常事態委員会は19日、ゴルバチョフ大統領を解任した」と、名詞句「大統領の解任」と、照応詞「この」と、空主語「 ϕ SUBJ」との間で照応が成立している。

テキスト1 ソ連の国家非常事態委員会は19日、ゴルバチョフ大統領を解任したと発表した。大統領の解任が西側の対ソ政策に重大な影響を及ぼすことは必至である。政府は、臨時の閣議を開き、この事態への対応を協議している。また、 ϕ SUBJ が替相場へ及ぼす影響も懸念されている。

テキスト1でこのような照応解釈が成立するのは、陳述縮約パラダイム (Jelinek et al. 1995) によれば、例えば文「ソ連の国家非常事態委員会は19日、ゴルバチョフ大統領を解任した」を X とし、名詞句「大統領の解任」を Y としたとき、これらが次の三つの制約を満たすからである。ここで、解釈 X と Y は構文木上の節点 X と Y にそれぞれ対応するものとする。

構文制約 Y はある構文木上で X の後方に位置する。

陳述縮約制約 Y は X を縮約した言語形式である。すなわち、 X と Y の陳述縮約度の間で表5に示す関係が成立する。

意味制約 Y の意味は X の意味と矛盾しない。すなわち、 X の人称、性、数が Y の人称、性、数にそれぞれ一致し、かつ、 X の意味標識と Y の意味標識が上位下位関係にある。

構文制約は比較的緩い制約であり、文間での前方照応と文内での前方照応を区別せずに扱える。このような統一的な扱いを可能にするために、入力が複数の文から成るテキストの場合でも入力全体を覆う構文木を生成する構文解析規則を図1の(a)と(b)のように記述している。

陳述縮約度は、ある言語表現が他の言語表現をどの程度指しやすいかと、他の言語表現からの程度指されやすさを表す。完全文の陳述縮約度を1、名詞句の陳述縮約度を7、日本語で表現する必要はないが英語では表現する必要のある照応詞の陳述縮約度を9のように定める。表5は、例えば節点 X の陳述縮約度が9であるとき、節点 Y の陳述縮約度は8か9でなければならないことを表す。構文木上の終端節点の陳述縮約度は辞書で与え、非終端節点の陳述縮約度は構文解析規則で与えている。

意味制約が満たされるかどうかの判定では、節点 X と Y の両方が終端節点である場合には、辞書に記述されているそれぞれの意味標識を照合すればよい。他方、節点 X と Y の両方あるいは一方が非終端節点である場合には、その木構造とそれを構成する終端節点の意味標識などに基づいて全体の意味を求めなければならないが、この処理は実現できていない。現在のところ、非終端節点の意味標識は、その主辞である終端節点の意味標識を特化した意味標識であるとしている。例えば動詞「解任した」の意味標識が AC(action) であるとき、「解任した」を主辞とする文「ソ連の国家非常事態委員会が19日、ゴルバチョフ大統領を解任した」全体が持つ

表 5 陳述縮約度に関する制約

節点 X の陳述縮約度	節点 Y の陳述縮約度
0	1, 2, 3, 4, 5, 6, 7, 8, 9
1	2, 3, 4, 5, 6, 7, 8, 9
2	3, 4, 5, 6, 7, 8, 9
3	4, 5, 6, 7, 8, 9
4	5, 6, 7, 8, 9
5	6, 7, 8, 9
6	7, 8, 9
7	7, 8, 9
8	8, 9
9	8, 9

意味標識を $AC+\alpha$ と表す。複合意味標識 $AC+\alpha$ は、原子意味標識 AC に何らかの意味 α が加わった AC の下位範疇を意味する。ここで、原子意味標識 A が原子意味標識 B の下位範疇であるとき、複合意味標識 $A+\alpha$ は B の下位範疇であると定める。また、 A が複合意味標識 $B+\beta$ の下位範疇であるかどうかと、 $A+\alpha$ が $B+\beta$ の下位標識であるかどうかは不明であると定める。二つの節点 X と Y の意味標識間の上位下位関係が不明である場合も、 X と Y は意味制約を満たすものとする。

節点 X と Y が構文、意味、陳述縮約に関する制約をすべて満たしているとき、二つの節点を Y から X へ向かうリンクで結ぶ。リンクには次の基準に従って照応に関する優先度を与える。

- (1) X と Y の意味標識の間に上位下位関係にあるかどうかは不明である場合、優先度を 0 とする。
- (2) X と Y の意味標識の間に上位下位関係にあることがわかっており、かつ Y の陳述縮約度が 7 以下である場合、優先度は 1 とする。
- (3) X と Y の意味標識の間に上位下位関係にあることがわかっており、かつ Y の陳述縮約度が 8 以上である場合、優先度を 2 とする。

テキストの照応解釈の優先度は、その解釈を構成するリンクの優先度の和であり、その値が最も高い解釈が最良の照応解釈である。テキスト 1 の場合、照応関係にある各言語表現は、図 4 に示すリンクで結ばれる。図 4 では、名詞句「大統領の解任」から文「ソ連の国家非常事態委員会は 19 日、ゴルバチョフ大統領を解任した」へのリンクの優先度が 0 であり、それ以外のリンクの優先度はすべて 2 であるので、この照応解釈の優先度は 10 となる。

照応関係にある各言語表現の間で人称、性、数、意味標識を伝播することによって、これらの曖昧性が解消される。図 4 の空主語「 ϕ_{SUBJ} 」は、人称、性、数、意味標識が三人称 (3rd)、中性 (n)、単数 (sg)、 $AC+\alpha$ となるので、代名詞 “it” と訳される。

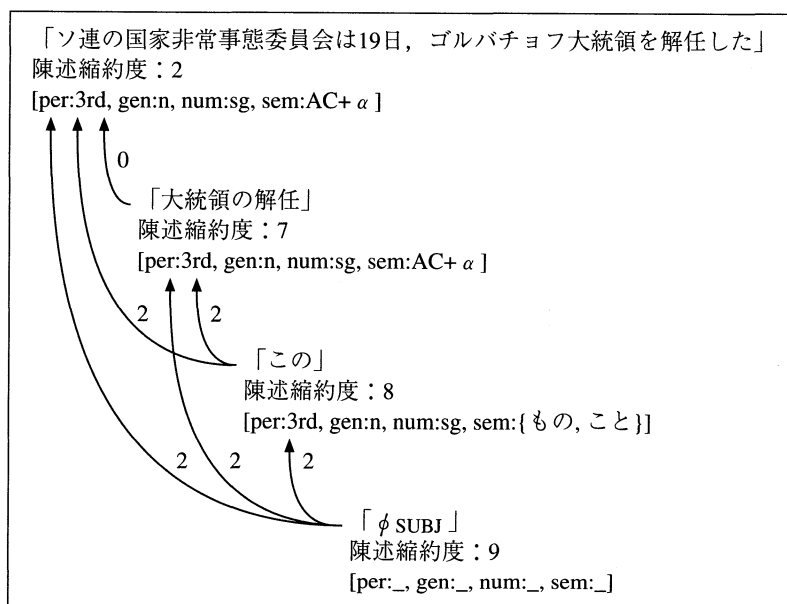


図 4 テキスト 1 における照応

6.3 構文，共起的意味，照応に関する優先度に基づく総合評価

共起的意味・照応解析系では，さらに，次のような総合評価式に基づいて構文，共起的意味，照応に関する各優先度を組み合わせた評価を行ない，解析系全体での (準) 最良解釈を選び出す⁵。

$$S = W_{\text{SYN}} \times S_{\text{SYN}} + W_{\text{SEM}} \times S_{\text{SEM}} + W_{\text{COR}} \times S_{\text{COR}} \quad (2)$$

S_{SYN} , S_{SEM} , S_{COR} はそれぞれ構文，共起的意味，照応に関する優先度であり， W_{SYN} , W_{SEM} , W_{COR} は各優先度についての相対的重要度である。

理想的なテキストでは，構文，共起的意味，照応の各最良解釈から全体での最良解釈が構成される可能性が高いと考えられる。これに対して現実のテキストでは，各優先度に基づく最良解釈の間で競合が生じ，各最良解釈が相容れないことがある。このような場合，どの解釈を優先させるかは，主に相対的重要度 W_{SYN} , W_{SEM} , W_{COR} を調整することによって経験的に決定する。今回の実験では， W_{SYN} , W_{SEM} , W_{COR} を，訓練用テキストの分析結果に基づいて，それぞれ 21, 3, 1 とした。

⁵ 形態素解析の精度は十分高いため，形態素に関する優先度と他の優先度との相互作用は考慮しない。

7 動詞型に基づく変換

動詞型に基づく変換の目的は、補足語に前置詞を付加することと、英文での補足語の位置を決定することである。この処理は図5に示すような規則に従って行なう。図5の手続き prefix() は、第一引数の補足語に、第二引数で指定された前置詞を付加する。手続き move() は、補足語を動詞の後方に移し、引数で指定された順序に並べ換える。引数に記述されていない補足語は元の位置に残しておく。例えば「少年が (SUBJ) 犬と (CP) 川を (PTR) 泳いでいる (vh).」という文から生成される構文木は、図5の規則により “boy swim {over|across|through} river with dog” に対応する木に変換される。

```

if(head_verb(vh)){
  prefix(AI, {from|by|with});
  prefix(CP, with);
  prefix(LT, {as far as|up to});
  prefix(ML, by);
  prefix(PTR, {over|across|through});
  prefix(TG, {for|to});
  move([PTR, CP, TG, PAC, AI, LT, ML]);
}

```

図5 動詞型に基づく変換規則

8 個別の変換情報に基づく変換

個別の変換情報による変換は、辞書の対訳情報に基づいて木構造を書き換える。対訳情報は英単語、中間記号、セパレータから成る。中間記号は、さらに、ここで扱う変換記号と生成系で扱う生成記号に分けられる。対訳情報の例を表6に挙げる。表6において、英大文字と数字は中間記号を、英小文字は英単語をそれぞれ意味し、それ以外はセパレータである。

表6 対訳情報

見出し表現	対訳情報
からの	B_from_A
たことがな	have_never_DONE
中に	AV2(while_SENT)
預け	entrust_P(#TG:0)_P(#OBJ:with)
容疑がかか	be_^_under_suspicion
動作	{function operate run}

変換記号は全部で 53 種類定義されているが、その一部を表 7 に示す。簡単な変換の例として「この理論は発話行為の観点からの意図が扱える。」という文に対する処理を図 6 に示す⁶。変換の前処理として、対訳情報を構造化し、変換記号なども木構造上の一節点とみなす。図 6 に現れている変換記号は名詞句の移動に関する A と B と、動詞の移動に関する DO であり、それぞれの移動を矢印で示すように行なう。なお記号[^], AT, T, ZR は生成記号である。

表 7 変換記号

変換記号	説明
A	構文木上でこの記号に最も近い前方の名詞句をこの記号の位置に移動する。
B	最も近い後方名詞句をこの位置に移動する。
DO, DID, DOING, DONE	最も近い前方の動詞をこの位置に移動する。さらに DID, DOING, DONE の場合はそれぞれ過去形, 現在分詞形, 過去分詞形を生成系で生成する。
#AG, #OBJ, #TG など.	それぞれの記号と同一節内に存在する AG, OBJ, TG などをもこの記号の位置に移動する。これによって、動詞型に基づく変換で決定された補足語の順序を変更する。
SENT, UOSENT, USSENT など.	最も近い前方の埋め込み節をこの位置に移動する。ただし、UOSENT, USSENT の場合はそれぞれ、SUBJ が存在しない節、OBJ が存在しない節でなければならない。
P	補足語に付加すべき前置詞を指定する。これによって、動詞型に基づく変換で決定された前置詞を変更する。

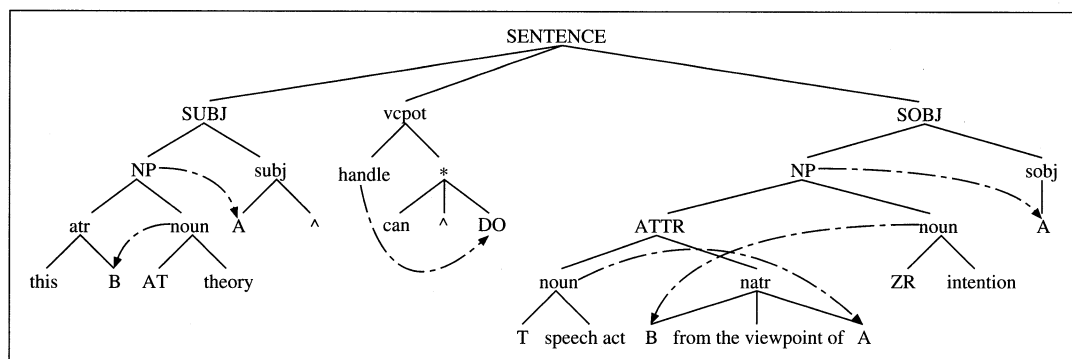


図 6 個別の変換情報に基づく変換の例

⁶ 図 6 は実際の木構造を単純化したものである。

9 生成

生成規則には、名詞句へ付加する冠詞を決定する規則、屈折形や派生形を決定する規則、副詞の最終的な位置を決定する規則、時制の調整を行なう規則などがある。これらのうち冠詞の生成に関しては、辞書で冠詞が一意に指定されている場合と文脈に依存して決まる場合があるが、後者の場合には照応解析の結果に基づいて冠詞を決定する。すなわち、ある名詞句が他の事象を指していれば定冠詞を選択し、そうでなければ不定冠詞あるいはゼロ冠詞を選択する。また、名詞の屈折形の決定に関しても、照応解析結果に基づいて単数形、複数形の生成を行なう。

10 翻訳実験

10.1 実験方法

実験には、池原らによって編集された機械翻訳機能試験文集(池原, 白井, 小倉 1994) の 2868 文を用いた。まず、試験文集に合わせた準備を全く行なわない完全ブラインドテストを行ない、その評価結果に基づいて辞書と規則を修正した後、ウィンドウテストを行なった。試験文集だけに合わせた修正は極力避け、一般性のある修正を行なうように努めた。翻訳品質の評価、辞書と規則の修正、試験のサイクルは四回繰り返した。評価および修正はすべて一名で行なった。完全ブラインドテストの結果の評価から第四回目のウィンドウテストの結果の評価までに要した期間は、およそ六ヶ月であった。

翻訳品質評価の基準は、訳文が 1) 文法的か、2) わかりやすいか、3) 原文の意味と一致するか、4) 利用者の役に立つかという観点から表 8 のように設定した。この評価基準において、4 点以上を合格とし、それ未満を不合格とする。

10.2 実験結果

完全ブラインドテストと四回目のウィンドウテストでの評価点の分布を表 9 に示す。合格した文の数は、完全ブラインドテストでは全体の 46.4% にあたる 1332 文、ウィンドウテストでは 2096 文 (73.1%) であった。評価点の平均は、完全ブラインドテストでは 2.7 点で合格点に達しなかったが、ウィンドウテストでは 4.2 点と合格点を上回った。各評価点を与えられた文の数が完全ブラインドテストとウィンドウテストでどのように変化したかを見ると、合格領域の 6 点ないし 4 点となる文数はいずれも増加し、不合格領域の 3 点ないし -2 点となる文数はいずれも減少している。特に、何も出力されないために 0 点となる文数の減少が著しい。これは主に、完全ブラインドテストでは入力文全体を覆う構文木が生成できず何も出力されなかった文に対して、ウィンドウテストでは全体の構文木が生成できるようになったことによる。

各文の評価値が完全ブラインドテストとウィンドウテストでどのように変化したかの分布を

表 8 翻訳品質評価の基準

評価点	文法性	理解容易性	意味等価性	有用性
6	文法的である.	容易に分かる.	一致する.	役立つ.
5	文法的である.	注意深く読めば分かる.	ある文脈では一致する.	対象分野や話題に関する知識を持つ利用者には役立つ.
4	文法的だが不自然である.	注意深く読めば分かる.	ある文脈では一致する.	同上.
3	非文法的だが原語がすべて訳された.	注意深く読めば分かる.	ある文脈では一致する.	同上.
2	非文法的で原語が残っている.	注意深く読めばほぼ分かる.	ある文脈ではほぼ一致する.	日本語辞書で単語を探す能力を持つ利用者には役立つ.
1	非文法的で原文を参照する必要がある.	注意深く読めば部分的に分かる.	ある文脈では原文と矛盾しない.	両言語の知識が無いと役に立たない
0	何も出力されない.	分からない.	どんな文脈でも原文と一致しない.	役に立たない.
-1	必須情報が欠落している.	分からない.	どんな文脈でも原文と一致しない.	役に立たない. 理解してみる努力が無駄になるので原文を読んで理解してみるほうがよい.
-2	文法的である.	分かる.	どんな文脈でも原文と一致しない.	危険. 訳文の質が高いほどより危険である. プロの翻訳者でさえ訳文を信じてしまい原文と照らし合わせない恐れがある.

表 10 に示す. 表 10 によれば, 評価点が向上した文数 (表の左下隅の領域) は 1186 文 (41.4%) であり, そのうち不合格から合格へ改善された文数は 835 文である. 逆に, 評価値が低下した文数 (表の右上隅の領域) は 204 文 (7.1%) であり, そのうち合格から不合格へ悪化した文数は 71 文である. また, 両テストで評価点に変化がなかった文数 (表の対角線上) は 1478 文 (51.5%) である.

この結果から, 辞書と規則の修正による悪影響を比較的小さく抑えつつ, 翻訳品質の改善が実現できているといえる.

完全ブラインドテストとウィンドウテストでの翻訳例を付録の表 11 に示す.

11 おわりに

本稿では, 日英機械翻訳システム TWINTRAN の辞書と規則について述べ, NTT 機械翻訳機能試験文集を対象として行なった翻訳品質評価実験の結果を示した. ウィンドウテストでは,

表 9 ブラインドテストとウィンドウテストでの評価点の分布

評価点	ブラインド	ウィンドウ	増減文数
6 点	645 文(22.5%)	1054 文(36.7%)	409
5 点	399 文(13.9%)	616 文(21.5%)	217
4 点	288 文(10.0%)	426 文(14.9%)	138
3 点	389 文(13.6%)	304 文(10.6%)	-85
2 点	108 文 (3.8%)	106 文 (3.7%)	-2
1 点	41 文 (1.4%)	2 文 (0.1%)	-39
0 点	434 文(15.1%)	95 文 (3.3%)	-339
-1 点	439 文(15.3%)	221 文 (7.7%)	-218
-2 点	125 文 (4.4%)	44 文 (1.5%)	-81

表 10 ブラインドテストとウィンドウテストでの評価点の変化

ブ \ ウ	6 点	5 点	4 点	3 点	2 点	1 点	0 点	-1 点	-2 点
6 点	595	26	9	4	3	0	4	3	1
5 点	59	304	15	8	2	0	8	3	0
4 点	46	36	171	20	4	0	2	9	0
3 点	85	65	65	146	12	0	5	11	0
2 点	19	13	23	17	27	0	2	6	1
1 点	12	12	10	3	0	1	1	2	0
0 点	98	94	81	50	17	0	59	33	2
-1 点	98	56	42	49	33	1	9	143	8
-2 点	42	10	10	7	8	0	5	11	32

我々の評価基準で、試験文集の 73.1% の文が合格となり、試験文集全体の平均点も合格点を上回る結果が得られた。

今回は、複数の文から成るテキストを対象とした評価実験は行なわなかったため、文間照応の認識によって得られる効果が確認できなかった。今後、テキストを対象とした実験を行なう必要がある。

謝辞

形態素解析系の設計と開発を行なって頂いた Graham Wilcock 氏 (現在 UMIST. 当時シャープ (株)) に感謝します。ALT-J/E に関する文献を提供いただいた Francis Bond 氏 (NTT) と、試験文集を作成された関係者の方々に感謝します。さらに、有益なコメントを頂いた査読者の方に感謝します。

参考文献

Den, Y. (1996). *A Uniform Approach to Spoken Language Analysis*. Doctoral Dissertation, School of Electrical and Electronic Engineering, Faculty of Engineering, Kyoto Univer-

sity.

- Ferrari, S. (1996). "Using Textual Clues to Improve Metaphor Processing." In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL96)*, pp. 351–353. Also published as <http://xxx.lanl.gov/abs/cmp-lg/9607034>.
- Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. English Language Series 9. Longman.
- Hobbs, J. R. and Bear, J. (1990). "Two Principles of Parse Preference." In *Proceedings of the 13th International Conference on Computational Linguistics (COLING90)*, Vol. 3, pp. 162–167.
- Ikehara, S., Shirai, S., and Bond, F. (1996). "Approaches to Disambiguation in ALT-J/E." In *Proceedings of the International Seminar on Multimodal Interactive Disambiguation (MIDDIM96)*, pp. 101–117.
- 池原悟, 白井諭, 小倉健太郎 (1994). "言語表現体系の違いに着目した日英機械翻訳機能試験項目の構成." 人工知能学会誌, **9** (4), 569–579. <http://www.kecl.ntt.co.jp/icl/mtg/resources/mt-test-set-1.txt>.
- Jelinek, J., Yoshimi, T., Nishida, O., Tamura, N., and Murakami, H. (1995). "Text-Wide MT Grammar." In *Proceedings of the 3rd Natural Language Processing Pacific Rim Symposium (NLPRS95)*, pp. 449–454.
- Kay, M. (1980). "Algorithm Schemata and Data Structures in Syntactic Processing." Technical Report CSL-80-12, Xerox PARC.
- 長尾確, 丸山宏 (1992). "自然言語処理における曖昧さとその解消." 情報処理, **33** (7), 746–756.
- 中岩浩巳, F. Bond (1997). "日英機械翻訳における文脈処理技術." *NTT R&D*, **46** (12), 1411–1418.
- 柴田勝征 (1996). "US 式和英翻訳システムの製作." <http://www1.sm.fukuoka-u.ac.jp/~kshibata/chosho/waei.j.html>.
- 柴田勝征 (1998). "US 式翻訳システムの特徴再論—最近の機械翻訳研究者の問題意識に即して—." テクニカル・レポート, 「言語・認識・表現」研究会. http://www1.sm.fukuoka-u.ac.jp/~kshibata/chosho/lace98_8.html.
- 島津明, 内藤昭三 (1989). "日本語の読みの偏好のモデル化." 研究報告 NL73-9, 情報処理学会.
- Tamura, N., Bos, M. J. W., Murakami, H., Nishida, O., Yoshimi, T., and Jelinek, J. (1991). "Lazy Evaluation of Preference on a Packed Shared Forest without Unpacking." In Brown, C. G. and Koch, G. (Eds.), *Natural Language Understanding and Logic Programming III*, pp. 13–26. Elsevier Science Publishers B.V.
- 田中穂積 (1989). 自然言語解析の基礎. 産業図書.

- Tsujii, J., Muto, Y., Ikeda, Y., and Nagao, M. (1988). "How to Get Preferred Readings in Natural Language Analysis." In *Proceedings of the 12th International Conference on Computational Linguistics (COLING88)*, Vol. 2, pp. 683–687.
- Wilks, Y. (1978). "Making Preferences More Active." *Artificial Intelligence*, **11**, 197–223.
- 吉見毅彦, Jiri Jelinek, 西田収, 田村直之, 村上温夫 (1997). "Text-Wide Grammar に基づくテキスト解析." 自然言語処理, **4** (1), 3–21.

略歴

吉見 毅彦: 1987 年電気通信大学大学院計算機科学専攻修士課程修了。1987 年よりシャープ (株) にて機械翻訳システムの研究開発に従事。1999 年神戸大学大学院自然科学研究科博士課程修了。

Jiri Jelinek: チェコのプラハの Universita Karlova 卒業 (言語学・英語学・日本語学)。1959 年以来, 日英機械翻訳実験中。英国 Sheffield 大学日本研究所専任講師を 1995 年退職。1992 年より 1996 年までシャープ専任研究員。

西田 収: 1984 年大阪教育大学教育学部中学校課程数学科卒業, 同年より神戸大学工学部応用数学科の教務補佐員として勤務。1987 年シャープ (株) に入社。現在, 同社情報家電開発本部 NB 第一プロジェクトチームに所属。情報処理学会会員。

田村 直之: 1985 年神戸大学大学院自然科学研究科システム科学専攻博士課程修了。学術博士。同年, 日本アイ・ビー・エム (株) に入社し東京基礎研究所に勤務。1988 年神戸大学工学部システム工学科助手。講師を経て, 現在同大学工学部情報知能工学科助教授。論理型プログラミング言語, 線形論理などに興味を持つ。

村上 温夫: 1952 年大阪大学理学部数学科卒業。神戸大学理学部助手, 講師, 教養部助教授を経て, 1968 年より工学部教授。この間, University of Kansas 客員助教授, University of New South Wales 客員教授, Nanyang University 客員教授を併任。1992 年より 1998 年まで甲南大学理学部教授。神戸大学名誉教授。理学博士 (東京大学)。関数解析, 偏微分方程式, 人工知能, 数学教育などに興味を持つ。著書に "Mathematical Education for Engineering Students" (Cambridge University Press) など。日本数学会, 日本数学教育学会各会員。

(2000 年 1 月 24 日 受付)

(2000 年 3 月 9 日 再受付)

(2000 年 6 月 30 日 採録)

付録

表 11 翻訳例

原文	訳文	評価点
独自に開発したソフトを組み込むことで、接続可能台数を増やした。	(ブ) It multiplied the number of connectable devices by incorporating the software which it uniquely developed.	6
	(ウ) It multiplied the number of connectable devices by incorporating the software which we uniquely developed.	6
その問題は調査するのに、時間を要する。	(ブ) Although it surveys that problem, it demands time.	-2
	(ウ) We require time to survey that problem.	6
私はそばよりうどんが食べたい。	(ブ) I want to eat udon from the side.	-2
	(ウ) I want to eat udon from the side.	-2
下の部屋は誰もいない。	(ブ) The room below has not anyone.	6
	(ウ) There is no room below.	-2
行った先は雨だった。	(ブ) Before it went was the rain.	-2
	(ウ) Before we went was the rain.	-2
人数が制限され、料金が安いときのみ、私は参加する。	(ブ) The number is limited and only, I participate when a fee is cheap.	3
	(ウ) The number is limited and I participate only when a fee is cheap.	5
彼はライバルと差が開いてしまった。	(ブ) He ended up increasingly differing as the rival.	3
	(ウ) He ended up establishing a lead over a rival.	6
計算機の製造には半導体工場などが必要だが、中国ではこうしたハイテク設備の建設技術をもつ技術者の絶対数が不足している。	(ブ) ### 構文解析失敗のため出力なし ###	0
	(ウ) Expert's in technology the semiconductor plant etc is required in production and who have building this kind of Hi-Tech equipment technology in China of computers absolute number is falling short.	-1
相手の会社が契約の期間までに必要なすべてのデータを揃える事ができなかった場合、当方に対して納得いく理由を示さない限り、当社はそのデータの受領を拒否し、損害の保証を求めることができる。	(ブ) ### 構文解析失敗のため出力なし ###	0
	(ウ) If partner's company could not assemble all data necessary even for terms of contracts against this side, as far as it does not indicate the sufficient reason, our company denies the acceptance of those data and you can demand guaranteeing damage.	4